 **Paper/Code:** Econometrics -MDS531A

Econometrics Labs Report  
(1-6)

**Prepared by**

Utkarsh Misra **{2448370}**

**Faculty Incharge**

Dr. Anusha James

**Lab 1**

**Education and Earnings: An Econometric Study with Logistic Regression**

**Objective:** This project's main purpose is to investigate education's effect on an individual’s income using cross-sectional economic data from the Adult Census Income dataset. More specifically, the objective is to use a logistic regression model to predict the likelihood of an individual earning over $50,000 in income (or more) based on years of education. This project will provide evidence on the relationship between education and income which serves as evidence of human capital as an economic factor.

**Dataset Description:** The dataset used in this research is the Adult Census Income Dataset, also known as the UCI Adult dataset. It consists of 32,561 records with 15 attributes, all of which are a combination of demographic, educational, occupational, and economic information gathered by the U.S. Census Bureau. The dataset aims to predict an individual's income level, which has been classified as being less than or equal to 50,000 USD (<=50K) and greater than 50,000 USD (>50K).

The dataset includes categorical and numerical variables. For the numerical variables, age, fnlwgt (sampling weight provided by the census), education\_num (numeric coding of education), capital\_gain, capital\_loss, and hours\_per\_week offer a quantifiable approach to personal and work-related information. The categorical variables: workclass, education, marital\_status, occupation, relationship, race, sex, and native\_country describe qualitative measures informally such as type of employment, educational achievement, family structure, race, sex, and country of origin. The income variable is categorical and binary as a target variable for classification.

**Results & Interpretation:** A logistic regression model was fitted with income category (1 if income > $50,000; 0 otherwise) as the dependant variable and years of education (education\_num) as the explanatory variable. The fitted equation is:

logit(P)=−5.0197+0.3643×education\_num

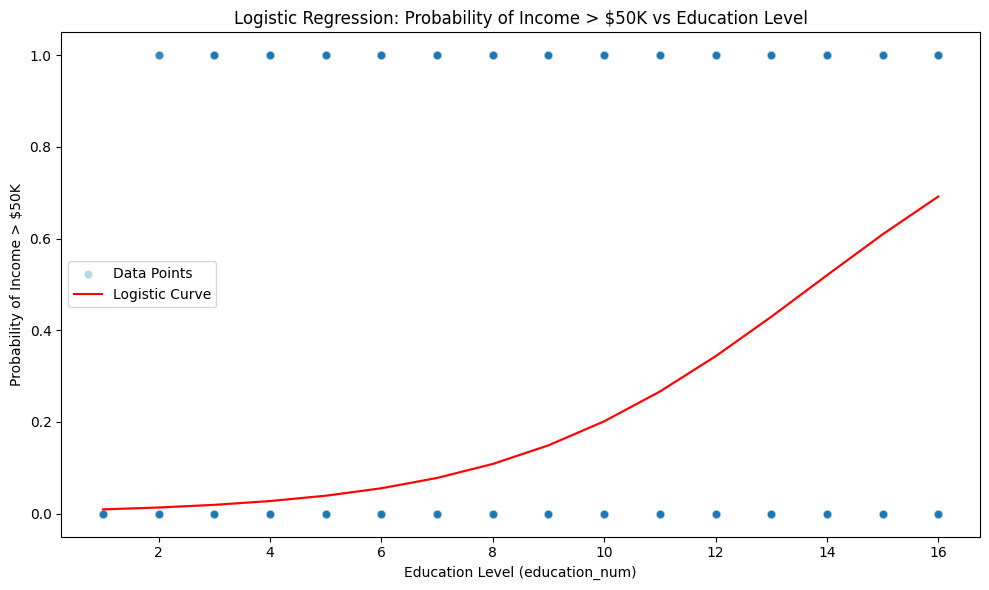
where P is the probability of earning more than $50,000 annually.

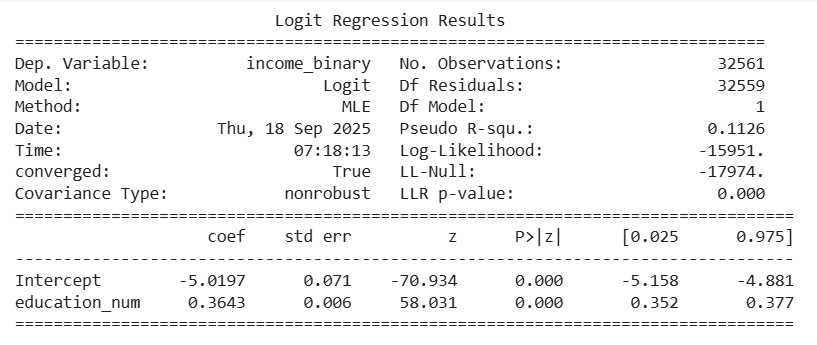
The coefficient for education\_num is 0.3643 and statistically significantly different from 0 at the 1% level (p-value < 0.001). The positive coefficient indicates a greater likelihood of making more than $50,000 with a higher number of years of education. The intercept is -5.0197, which is also statistically significant. The pseudo R² measure of McFadden’s 0.1126 indicates that the educational variable alone accounts for approximately 11% in the variance of the income categories. A likelihood ratio test definitively rejects the null that the coefficient is equal to 0 (LLR p-value = 0.000).

These results suggest strong and positive association between education and income. Each additional year of education increases the log-odds of making more than $50,000 by an additional 0.3643. To convert this to odds ratio:

Odds Ratio=e0.3643≈1.44

This indicates that each additional year of education increases odds of making more than $50,000 annually by roughly 44%, holding other variables constant. The logistic curve below shows this relationship. As you get further along in educational attainment, the probability of earning high income increases steeply, particularly beyond the mid-education levels.





**Conclusion**: The results support the conclusion that education is an important factor influencing income in this cross-section. Those with more education had much higher odds of earning the specific figure of above $50,000 as compared to those who had fewer years of education. The model explains a relatively small portion of explained variance in income (Pseudo R² = 0.1126), although it demonstrates an important role of human capital in economic outcomes. Future models may include additional variables, for example, work experience, occupation, and demographic variables to illustrate the determinants of income more completely.

**Lab 2**

**Application of ARIMA Models for Macroeconomic Forecasting: Evidence from India’s GDP**

**Objective:** The aim of this lab is to use time series econometric methods to study and predict an important macroeconomic time series of your selection from the World Bank Development Indicators (WDI) database.

**Dataset Description:** This research utilizes a dataset from the World Development Indicators (WDI) database, published by the World Bank, containing macroeconomic and development indicators across countries and years. The structure of the file follows the standard World Bank format, which begins with a few rows that provide meta data (i.e., source of data, update date, and details about those indicators) which is followed by the main, large dataset.

Each observation in the file is a country–indicator–year observation. The key fields which are included in the dataset are:

* **Country Name** – Name of the country that is reporting
* **Country Code** – The three-letter ISO country code
* **Indicator Name** – Name of the economic or social variable being measured (e.g., GDP; trade; energy use; population)
* **Indicator Code** – World Bank code assigned to the indicator
* **Yearly Values** – Annual time series of the indicated years for selected indicators (generally from 1960 and subsequently, according to availability by indicator and country)

For this lab, a subset of the indicators was used to focus on the underlying determinants of economic growth in the Southeast Asia region. The indicators of interest were:

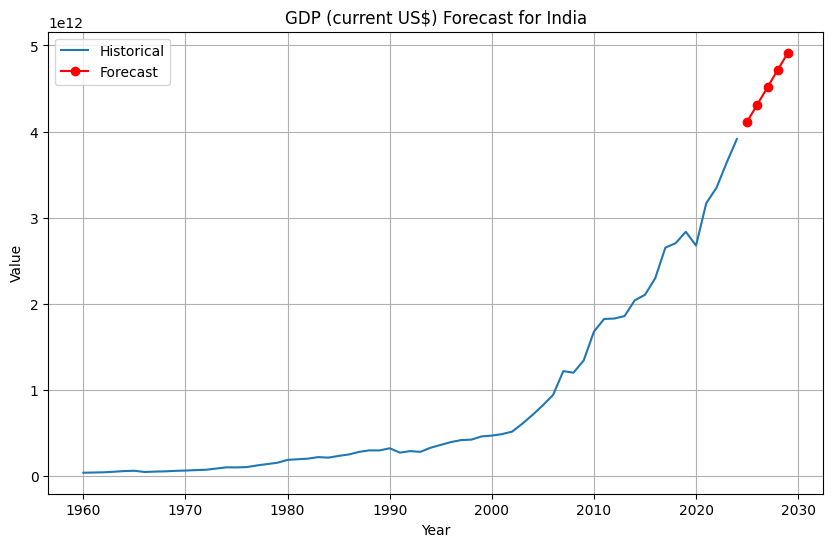
* **GDP (constant 2015 US$) –** Measure of economic output with adjustments for inflation
* **Energy use (kg of oil equivalent per capita) –** Captures industrial activity and energy use.
* **Trade (% of GDP) -** this shows the impact and integration with the global markets.
* **International tourism, number of arrivals -** this indicates the contribution of the tourism sector to the economy.

The data set is originally from 1990-2014 for five South East Asian countries: Vietnam, Indonesia, Malaysia, Thailand, and the Philippines. Once the data is cleaned and reshaped, it was converted into a balanced panel data set for econometric analysis through Fixed Effects and Random Effects models.

**Results & Interpretation:** Initially, the Augmented Dickey-Fuller (ADF) test was used to test for stationarity in India’s time series of GDP (current US$). The test produced an ADF statistic of 1.90 and a p-value of 0.9985, which is well above the 0.05 significance level. This indicates a non-stationary series, which will require differencing before it can be modeled using an ARIMA model.

After applying appropriate transformations, the series was fitted with a an ARIMA model. The historical series of GDP from 1960 to 2023 shows a stable upward trend with accelerated growth after the 1990s. The ARIMA forecast for the period 2024-2029 shows a continuation of this solid upward trend, forecasting India’s GDP to near 5 trillion US$ by 2029.

The red forecast line appears to reconcile with the historical series, indicating that the model fits the covariance dynamics of the growth. However, with the high degree of non-stationarity of the original data, it is likely that forecasts represent the persistence of the growth and not cyclical changes



**Conclusion**: The forecasts produced by the ARIMA-based forecasting exercise suggest that India's GDP (current US$) is strongly non-stationary in nature and appears to be following an exponential growth path. The forecasts project that continued economic growth is likely to occur, with GDP exceeding 4 trillion US$ in 2025 and moving toward the level of 5 trillion US$ by 2029, in line with the prevailing views and expectations among policymakers and market participants of India as one of the largest economies in the world.

While the ARIMA model can be relied upon to deliver reasonably short-term forecasts, the nature of the model places a reliance on past trends, meaning it would be incapable of capturing structural changes (example: global recessions, policy restructuring, pandemics etc.). In the future, a more accurate modelling of forecasts may be achieved with exogenous variables (ARIMAX) including investment, trade, or energy consumption, or other sophisticated machine learning forecasting techniques.

**Lab 3**

**Econometric Investigation of Macroeconomic Determinants of Growth: Evidence from Southeast Asian Countries**

**Objective:** The main goal of this lab is to utilize econometric techniques to examine the connection between economic growth and major macroeconomic variables for a selected group of Southeast Asian countries.

**Dataset Description:** The data for this lab is derived from the World Bank Development Indicators. It is in a panel dataset format, which implies that it incorporates both cross-sectional and time series dimensions. The data commenced in 1990, and covers the time-frame of 1990–2014 for selected Southeast Asian countries, namely, Vietnam, Indonesia, Malaysia, Thailand

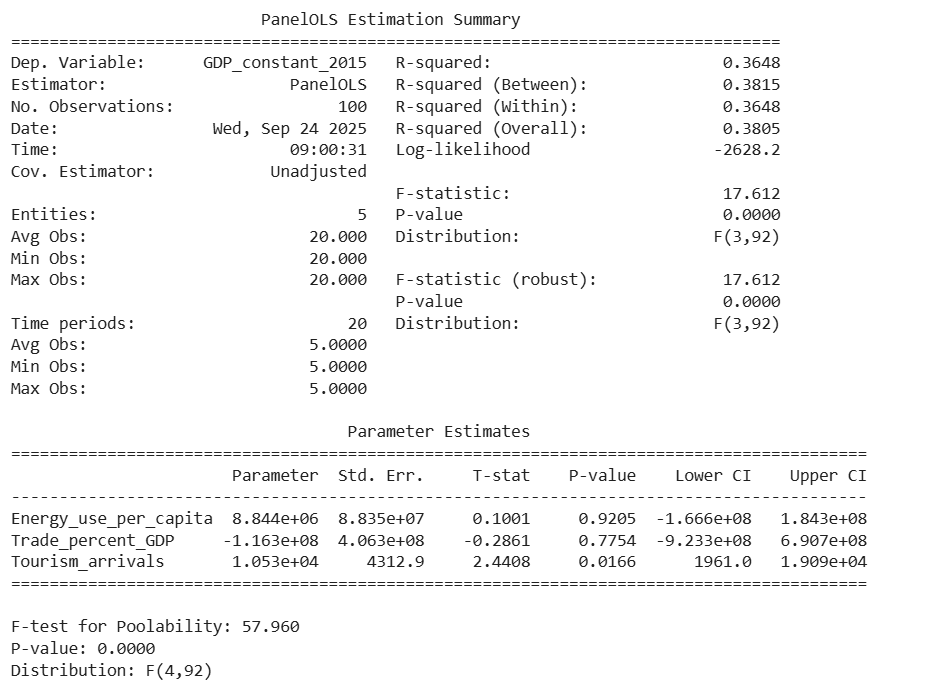
The dataset contains four major macroeconomic indicators. GDP (constant 2015 US$): a measure of economic production adjusted for inflation. Energy use (kg of oil equivalent per capita): reflects energy use and industrial activity. Trade (% of GDP): measures the degree of trade openness and country's integration into the global economy. International tourism, number of arrivals: measures the impacts of globalization and the contribution of the service sector.

The dataset has been reshaped from wide to long format so that it can be employed for econometric analysis. In its final form, each row consists of a unique country-year observation of a particular variable, enabling the ability to run panel regressions and time series analysis across countries.

**Result & Interpretation:** The panel regression model was estimated with the Fixed Effects (FE) specification to account for unobserved heterogeneity at the country level. The dependent variable is GDP (constant 2015 US dollars) and the independent variables are energy use (per capita), trade (% of GDP), and international tourism arrivals.

The model explains approximately 36% of the variation in GDP (R-squared = 0.3648 within). The overall F-statistic tests (17.61, p < 0.01) suggest that the independent variables are jointly significant.

* Energy Use (per capita): The coefficient is positive but statistically insignificant (p = 0.92). This indicates that, among these countries, variations in energy use per capita have no strong direct effect on GDP once country-specific effects are accounted for.
* Trade (% of GDP): The coefficient is negative and statistically insignificant (p = 0.78). This means that trade openness does not seem to generate statistically significant growth in GDP across any of the countries in the sample. This may be due to structural differences in trade balances or due to external shocks.
* Tourism Arrivals: The coefficient is positive (10,530) and is statistically significant at the 5% level (p = 0.0166). This suggests that an increase in international tourist arrivals is positively associated with GDP and reinforces the importance of the tourism sector as a driver of economic growth.



**Conclusion:** This econometric analysis demonstrates that international tourism plays a significant role in promoting GDP growth in Southeast Asian economies during the period 1990–2014. Conversely, energy consumption and trade openness were found to be statistically insignificant, suggesting that their effects may either operate in the long run or be influenced by structural differences among the countries.

The results imply that policy efforts to strengthen the tourism industry—such as improving infrastructure, connectivity, and service quality—could contribute to higher economic growth. At the same time, policymakers should recognize that trade and energy remain crucial sectors, but their contributions may require complementary policies (e.g., industrial diversification, sustainable energy use) to translate into measurable GDP gains.

**Lab 4**

**Time Series Regression Analysis of Unemployment Rate Using OLS and GLS**

**Objective:** This lab aims to discuss the U.S. monthly unemployment rate (UNRATE) through lagged regression, stationarity test, and Generalized Least Square (GLS) to account the autocorrelation, and to graphically and intuitively interpret the findings.

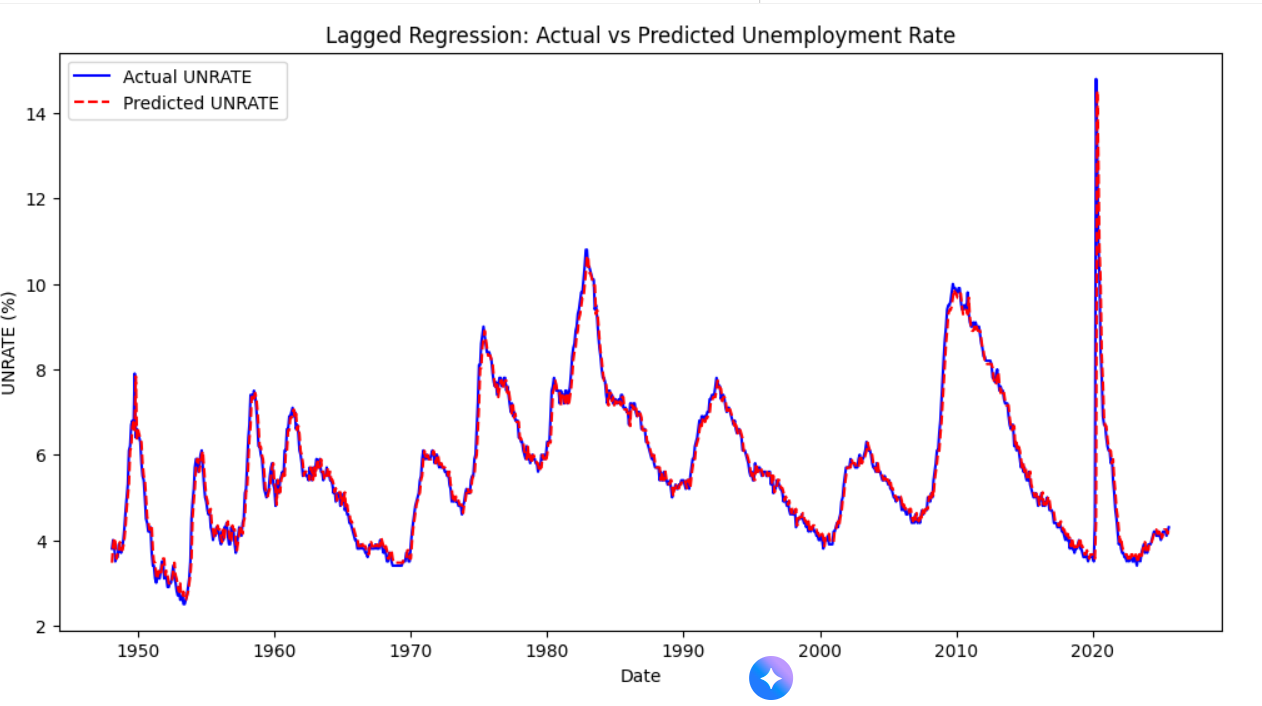
**Dataset Description:** The data involved in the present lab is the U.S unemployment rate (UNRATE), which is obtained via the Federal Reserve Economic Data (FRED). It is a monthly sequence of time series, recording the percentage of the working classes that are unemployed and actively looking to be employed. The dataset is usually a few decades long, which gives a long-term view on the dynamics of the labor market. It will be a two column table, where one column will be the date of observation in the format YYYY-MM-DD, and the other column will be the unemployment rate in percentage. The UNRATE series is not stationary at the level because of the trends and cyclic fluctuations, and is also autocorrelated because the rates of unemployment are not random over time. There are also structural changes like those that occur in times of economic recessions and hence it is a perfect dataset to study autoregressive models and also to test more complicated econometric models such as GLS. This data is applied across forecasting, economics and education to illustrate time series regression, stationarity test as well as the correction of autocorrelated residues.

**Results & Interpretation:** The OLS regression of the U.S. monthly unemployment rate (UNRATE) on its lagged value (UNRATE\_lag1) shows a very high R-squared value of 0.942, indicating that approximately 94% of the variation in the current unemployment rate is explained by the previous month’s rate. The coefficient of UNRATE\_lag1 is 0.9699, which is highly significant (p < 0.001), demonstrating strong persistence in unemployment rates over time. The constant term is small but significant (0.1719), indicating a slight baseline effect.

The Durbin-Watson statistic of 1.898 suggests minor positive autocorrelation in residuals. The residuals themselves are highly non-normal (extreme skewness and kurtosis), likely due to spikes during economic recessions. However, the ADF test on residuals confirms they are stationary, which validates the regression and indicates that the model is not spurious. Overall, the results confirm that unemployment rates are highly autocorrelated and can be effectively modeled using a lagged regression framework.

The Generalized Least Squares Autoregressive Model (GLSAR) regression of the U.S. unemployment rate (UNRATE) on its lagged value (UNRATE\_lag1), results in an R-squared value of 0.935, which slightly trails behind the OLS model (0.942), reinforcing that the model continues to explain a substantial amount of variation in UNRATE. The coefficient of UNRATE\_lag1 is 0.9670 and is highly statistically significant (p < 0.001), strengthening the presence of autocorrelation in the unemployment rates. The intercept measures at 0.1878, small but statistically significant for the GLSAR estimation.

Most importantly, the Durbin-Watson statistic measures at 1.997, very close to 2, indicating that the GLSAR has appropriately corrected for the autocorrelation exhibited from the OLS regression residuals. While the residuals display non-normality (high skew and kurtosis), it is expected because of economic shocks and recessions. GLSAR provides a more accurate estimate by correcting for the autocorrelation in the time-series, which makes inference stronger than the OLS regression.



**Conclusion:** The analysis shows that U.S. unemployment rate dynamics are highly persistent, with the previous month's unemployment rate serving as a useful predictor of the unemployment rate in the current month. The OLS lagged regression produces a statistically significant model, and the stationarity tests on the residuals confirm it. There is some minor autocorrelation, but the model adequately reflects the dynamics of unemployment. We can try to extend as a next step using Generalized Least Squares (GLS), which can adjust for autocorrelation and provide more efficient estimates. Overall, a lagged regression is reasonable for an econometric time series modeling and forecasting.

The findings substantiate the U.S. unemployment rate's high persistence over time, demonstrating that the lagged month is a significant predictor of the current month's unemployment rate. While OLS provides an acceptable first-stage model, GLSAR provides an improved model by adjusting for autocorrelation of the residuals, which is indicated by Durbin-Watson values closer to an ideal statistic of 2.00. Overall, both models indicate strong autoregressive behavior of unemployment, but GLSAR provides more efficient and consistent estimates that are better suited for econometric analysis. This validates the utilization of GLS in time-series models with autocorrelated errors, justifying it as an appropriate method for forecasting and policy analysis in labor economics.

**Lab 5**

**Regression Analysis and Outlier Detection Using the Mroz Dataset**

**Objective**: The primary aims for this lab are:

To run a linear-regression analysis of the Mroz data set with the aim of investigating the relationship between a dependent variable (e.g., hours worked by married women) and selected independent variables (e.g., education, experience, age, spouse's wage).To calculate leverage values and Cook's distance to identify influential observations in the data set and to visualize and analyze outliers and influential points for their effects on the regression model.

**Dataset Description:** The Mroz dataset originates from T.A. Mroz's 1987 study on the labor supply of married women. It contains 753 cross-sectional observations with 22 variables capturing personal, family, and economic characteristics, including:

* **inlf:** Labor force participation indicator (1 if employed, 0 if not).
* **hours:** Number of hours worked by the woman in 1975.
* **educ:** Years of education completed by the woman.
* **exper:** Labor market experience in years.
* **age:** Age of the woman.
* **huswage:** Hourly wage of the husband.
* **motheduc, fatheduc:** Education levels of mother and father.
* **unem:** County-level unemployment rate.

The dataset is widely used in labor economics and econometrics to study labor supply decisions and the influence of socio-economic factors on women’s labor market participation.

**Results & Interpretation:** Though numerous outliers and leverage points were detected, it is worth investigating or removing them for a better fit.

**Regression Analysis:**

* The regression model predicts hours worked by married women (hours) against variables educ, exper, age, huswage, motheduc, fatheduc, and unem.
* R-squared = 0.206 indicates 20.6% of the variability in hours worked is explained by the independent variables selected. The explanatory power is moderate, which is normal for cross-section socio-economic data.

**Significant predictors:**

* **exper (experience):** highly significant and positive (coef=48.741; p<0.001)- as expected, more experience leads to increased hours worked.
* **age:** negative and significant (coef=-17.5685; p<0.001)-older women tend to work fewer hours.

Other variables, like educ, huswage, motheduc, fatheduc, and unem, are not statistically significant at a 5% level, although educ and huswage are borderline.

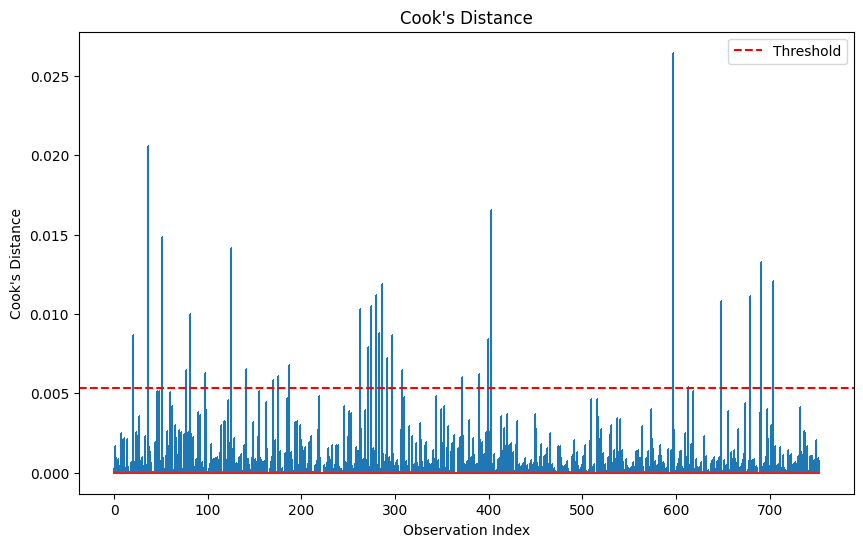
**Leverage and Cook’s Distance:**

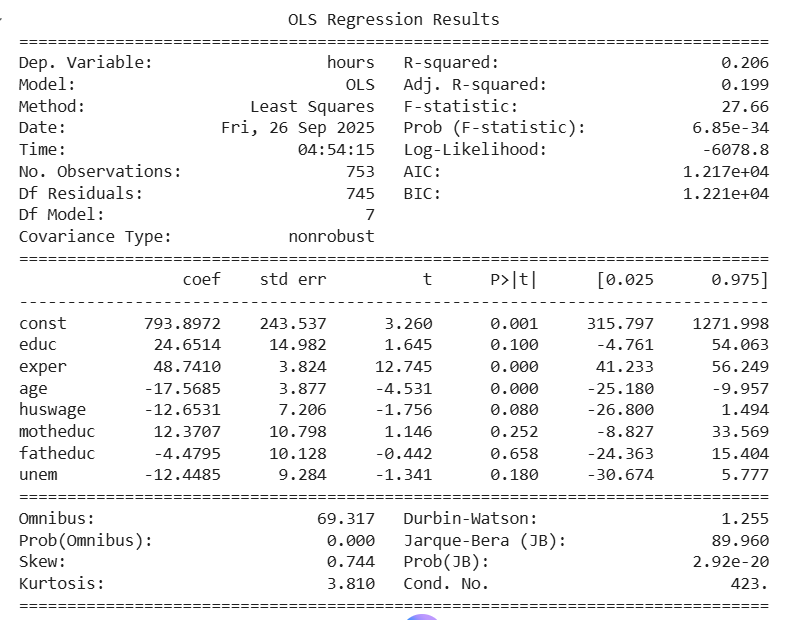
* High leverage points (indices: [31, 58, 77, 80, ...]) are observations that lie far away from the center of the independent variable’s space.
* High Cook’s distance points (indices: [20, 36, 51, 77, 81, ...]) are influential observations which are having large effect on the estimated regression coefficients.
* Observations 77, 81, 175, 286, 597, 679, and 703 were high leverage points and high Cook's distance; thus, making them particularly important, and should be given attention for further explorations.

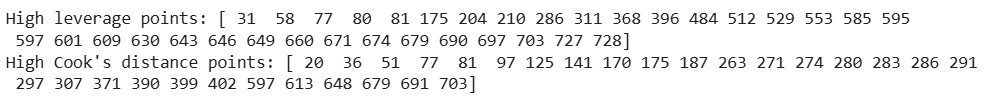
**Cook's Distance Plot:**

* From the plot, it is evident that most observations fall below the threshold line, while a few spikes are above the threshold; thus, there are influential points.
* These points might weaken the stability of coefficient estimates of the regression in case they are omitted.

The model clearly indicates that experience and age significantly impact women's hours of work, but informative evidence from influential observations must study the outliers and leverage points in the econometric analysis. Their removal or further exploration would certainly increase the reliability of the model.







**Conclusion:**

The regression model demonstrates the relationship of labor supply (hours) to socioeconomic factors in the Mroz dataset. Experience enhances hours while Age has a contrary effect. Other variables in the model are rather weak in explanation. Several observations are influential as per their leverage and Cook's distance. These points may unduly influence the regression coefficients and therefore require careful consideration. All in all, the lab illustrates the importance of integrating regression analysis with diagnostic measures such as leverage and Cook's distance so as to detect outliers and aid in model interpretation in econometrics.

**Lab 6**

**Analysis of 401(k) Participation Using OLS Regression: Addressing Omitted Variable Bias**

**Objective:**

To understand the concept of omitted variable bias in regression analysis and identify highly correlated independent variables and assess the effect of omitting them on the model. To fit an OLS regression model and evaluate its explanatory power using R²and perform the RESET test to check for model specification errors. To interpret the impact of omitting variables on coefficient estimates, significance, and overall model reliability.

**Dataset Description:**

The dataset used is 401KSUBS, sourced from Jeffrey Wooldridge’s *Introductory Econometrics: A Modern Approach*. It contains 9,275 observations of individuals’ financial, demographic, and retirement account information.

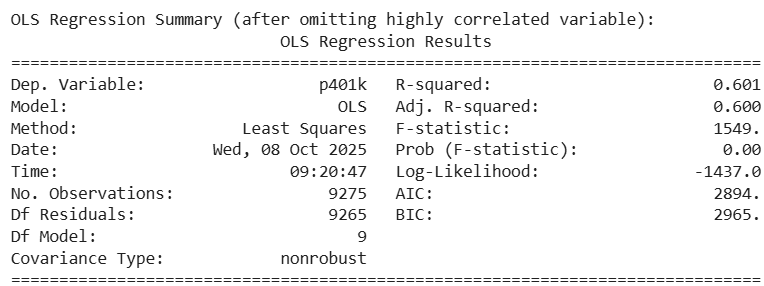
**Variables of interest:**

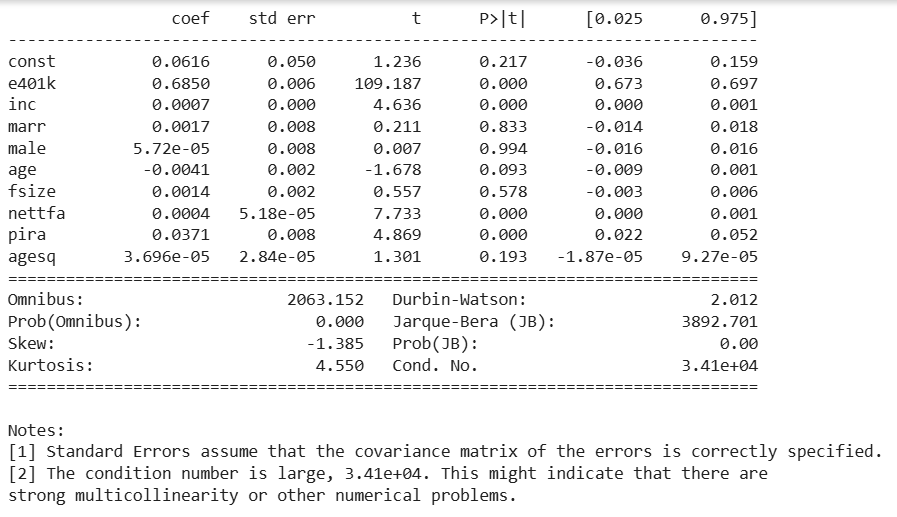
| **Variable** | **Description** |
| --- | --- |
| **e401k** | Indicator for 401(k) eligibility (1 = eligible, 0 = not) |
| **inc** | Annual income (in thousands of dollars) |
| **marr** | Marital status (1 = married, 0 = not) |
| **male** | Gender (1 = male, 0 = female) |
| **age** | Age of respondent |
| **fsize** | Family size |
| **nettfa** | Net total financial assets (in thousands of dollars) |
| **p401k** | Indicator for 401(k) participation (1 = participates, 0 = not) |
| **pira** | Indicator for having an IRA (1 = yes, 0 = no) |
| **incsq** | Square of annual income |
| **agesq** | Square of age |

For this analysis, we examined the correlation between independent variables and dropped incsq due to high correlation (0.94) with income to avoid multicollinearity.

**Result & Interpretation:**

* **Correlation Analysis:**
  + Income (inc) was highly correlated with its square (incsq) – 0.94.
  + Other variables had low to moderate correlation with inc (0.07–0.38), indicating minimal multicollinearity.
* **OLS Regression Results (after omitting incsq):**

****

****

* + **Model Fit:**
* R² = 0.601, meaning about 60% of the variation in 401(k) participation is explained by the model.
  + **RESET Test:**
* F-statistic = 132.51, p-value = 1.86e-30 → p-value < 0.05, so the model may be mis-specified, suggesting potential non-linear relationships or missing variables.

**Conclusion:**

401(k) eligibility (e401k), income (inc), net financial assets (nettfa), and IRA ownership (pira) are significant predictors of 401(k) participation. Other demographic variables like marital status, gender, age, family size, and squared age are not significant in this model. The model explains 60% of the variation in participation (R² = 0.601), which is reasonably good. The RESET test indicates possible model mis-specification, suggesting further investigation may be needed, such as including non-linear terms or interactions between variables.